"DATA MINING FOR BIOLOGICAL AND ENVIRONMENTAL PROBLEMS"

A Synopsis

Submitted in

In partial fulfillment for the award of the degree

Of

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

Supervised by
Dr. Manoj Shukla

Submitted by
JV'n Ms. Pooja Shrivastava

Department of Computer Science & Information Technology
Faculty of Engineering & Technology
Jayoti Vidyapeeth Women's University, Jaipur (Rajasthan)
March 2013

Introduction

Data mining (the superior scrutiny step of the "familiarity find in Databases" progression, or KDD), an interdisciplinary subfield of computer science[1] is the computational route of determine outlines in great data sets involving ways at the intersection of artificial intelligence, machine studying, statistics, and database systems. The in excess of aim of the data mining development is to remove in sequence from a data set and convert it into an comprehensible construction for more use. Aside from the raw investigation step, it involves database and data management aspects, data pre-processing, copy and conjecture deliberations, interestingness metrics, complexity deliberations, post-processing of discovered configurations, visualization, and online updating. The term is a exhortation, and is recurrently maltreatment to mean any form of bulky-scale data or in sequence processing (anthology, taking out, warehousing, breakdown, and information) but is also generalized to any class of computer conclusion support system, counting artificial intelligence, machine learning, and business intelligence. In the correct use of the word, the key expression is unearthing commonly defined as "detecting a bit new". Even the accepted book "Data mining: no-nonsense machine learning tools and methods with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the phrase "data mining" was simply added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to definite process, artificial intelligence and machine learning – are supplementary proper.

The definite data mining duty is the repeated or semi-automatic investigation of huge quantities of data to remove previously indefinite interesting outlines such as clusters of data proofs (cluster analysis), odd records (irregularity discovery) and dependencies (association rule mining). This frequently involves using database systems such as spatial indexes. These outlines can then be seen as a kind of précis of the input data, and may be used in further study or, for example, in machine learning and extrapolative analytics. For example, the data mining step power identify numerous groups in the data, which can then be used to obtain more accurate prediction results by a decision support system.

Background

Data mining idea and physical model from data has open for centuries. Early methods of identifying patterns in data comprise regression analysis (1800s) and Bayes' theorem (1700s). The proliferation, ubiquity and growing power of computer technology has improved strategy gift, thought, data group and storeroom. Data mining concepts build up more facility for study. As data sets have adult in size has more and more been increased with indirect, mechanical data processing, aided by other discoveries in information technology and computer science such as neural networks, cluster analysis, genetic algorithms (1950s).

Data mining researcher residential decision trees (1960s). Data mining researcher also urbanized support vector machines (1990s). Data mining is helpful to applying these methods. These methods are exposure hidden patterns in big data sets. Data mining bridges the gap from applied statistics, bioinformatics and artificial intelligence. Data mining provides the mathematical background to database management by exploiting. In Data mining data is stored and indexed in databases to implement the definite learning and sighting algorithms more ably. Data mining allowing such methods to be functional to a large of data sets.

What is Data Mining?

Data mining is a perfectly interdisplinary subject. Data mining can be explains in many different ways. Data mining should have been more unlike named first is knowledge mining from database and second is knowledge discovery from data set. Many people treat data mining and concepts as a synonym. It’s for another popularly used term and the others view data mining as simply an essential step in the course of data mining. The knowledge discovery process presents in figure 1.1 as an iterative run of the following steps. In this figure we can see that how data mining works.

Uses of Data mining

Games

Data mining is worn in games such as table cases with any foundation little-plank dots and boxes, small timber curse and positive end games in chess dots-and-boxes original are for data mining has been opened.

Business

Data mining is the scrutiny of chronological business. Data mining stored data as databases to reveal veiled patterns and trends. Data mining is very important in customer rapport management applications. In data mining data clustering is most dominant and important technique. It is also be used to repeatedly learn within a consumer list. Data mining is furthermore individual loyal to human property department in identifying the quality of their most thriving employees.

Science and engineering

In recent years, data mining has been worn in many areas of knowledge and industrial, such as bioinformatics, medicine, health care, physics and environment. In the study of human genetics, genetics, sequence mining helps tackle the very important goal of accepting. This thoughtful the mapping the inter personage variations in being DNA/RNA sequence and protein chain the unpredictability in disease openness. In simple terms, it aims to discover how the changes in an folks DNA/RNA sequence affects the risks of developing common diseases such as cancer, The data mining scheme is used to perform this task. This is known as multifactor dimensionality reduction.

Human rights

Data mining is practical in human rights area. Data mining of government records mostly of the courts system. Justice system enables the discovery of systemic human rights.

Objectives:

1. To Collect different samples from the different areas of the protein, and forest fire data.
2. To include identification and maintenance of techniques and explore clustering, classification techniques for this problem.
3. To perform comparative analysis between different areas (protein, and forest fire data).
4. To explore different techniques for protein, and forest fire data.
5. To use the data mining algorithms.
6. To explore the study of environmental data, Decision Tree, prediction data.

Review of Literature

In this literature (Agarwal et al., 1993) Proteins typically do not act remote in a cubicle but meaning with in complex cellular paths, interacting with additional proteins what's more in couples or as modules of superior complexes while lots of protein complexes have been known by great balance untried studies, due to a huge number of false-positive communications on hand in existing protein complexes, it is unmoving not easy to gain an correct considerate of purposeful modules, which cover assembly of proteins involved in general elementary biological occupation.

In this paper, we at hand a hyperclique example discovery come near for extracting purposeful modules (hyperclique patterns) from protein compound.

In this research discussed SAS solution for pharma covigilance.(Kumar et al., 2004),. (Arthur and Vassilvitskii, 2007) presented k-means++ method is a widely used clustering technique. Experiments show in this research argument improves both accuracy of k-means, and speed often quite dramatically.

In this literature (Basheer et al, 2012) authors presented Data mining. Data mining discover knowledge from a lot of quantity of data scrutiny. In this research deeply focused on, a genetic algorithm-bottom approach for mining arrangement set of laws. This research work based on coverage, comprehensibility of the rules, accuracy and simplifying the implementation of a genetic algorithm. In this research discussed the design of encoding, which is based on fitness function of genetic algorithm and genetic operators. Experimental results show that genetic algorithm proposed in this research. This is suitable for discovered higher classification performance and data mining methods to unknown data.

In (Bauer and Kohavi, 1999) presented a lot of examples bagging and boosting. In this research author presented MDT presented the better result voting and stacking. Bagging works as a method of increasing accuracy. Bagging is also called Bootstrap. Bootstrap is based on random sampling with replacement. Bagging is a machine learning ensemble algorithm.

In (Bifet et al., 2009) was present figures torrent construction. In this surround work author presented data mining framework which is applied with techniques.

In (Dietterich, 2002) described three types of ensembles problems associated with base learning algorithms: first is the statistical problem second is the computational problem, and the third is representational problem.

(Dr. Bhatia and Deepika Khurana, 2013) the authors presented k- Means clustering algorithm. This research presented the qualified study of customized k- Means clustering algorithm and original k-means clustering algorithm. Theses algorithms are execute on different datasets. Routine result came using Original k-Means and other adapted k-means algorithms. In this research used MATLAB R2009b. Authors the results are calculated on some performance measures such as, no. of points which is misclassified, accuracy, Silhouette validity index, no. of iterations and execution time in this research.

In (Hemalatha and Saranya, 2011) presented a survey paper of spatial data mining different spatial task. Authors focused on the sole features that distinguish spatial data mining from traditional data mining. In this research authors gave applications and Techniques, Issues and challenges on spatial data mining. Author presented spatial data analysis which is very different task for research area.

In (Dzeroski and Zenko, 2004) worked on Ensemble learning techniques. This technique created a meta-classifier by combining several classifiers, it is typically by voting, ensemble learning created on the same data. Ensemble learning improved their performance.

In this literature (Ester *et al*., 2001) authors presented a database-oriented scaffold for spatial data mining. A small set of basic and primary operations on these graphs and paths were discussed as database primitives for spatial data mining. Many types of techniques discussed in this research such as commercial DBMS were presented. In the research authors covered the main tasks of spatial data mining: spatial classification, spatial characterization, spatial clustering

and spatial trend detection. For each of these tasks, authors presented algorithms prototypical applications. The authors indicated interesting directions for future. Since the system overhead imposed by this DBMS is rather many types of concepts and techniques of improving the efficiency should be investigated. For example, techniques for processing sets which provide more information to the DBMS can be used to improve the overall efficiency of mining algorithms.

FCM is based on fuzzy logic (Elena, 2013) author presented paper is a survey of fuzzy logic theory which is applied in cluster analysis. In this work author presented review the Fuzzy c - means clustering method in MATLAB.

In this literature (Ester *et al*., 2001) authors presented a database-oriented scaffold for spatial data mining. A small set of basic and primary operations on these graphs and paths were discussed as database primitives for spatial data mining. Many types of techniques discussed in this research such as commercial DBMS were presented. In the research authors covered the main tasks of spatial data mining: spatial classification, spatial characterization, spatial clustering and spatial trend detection. For each of these tasks, authors presented algorithms prototypical applications. The authors indicated interesting directions for future. Since the system overhead imposed by this DBMS is rather many types of concepts and techniques of improving the efficiency should be investigated. For example, techniques for processing sets which provide more information to the DBMS can be used to improve the overall efficiency of mining algorithms.

(Ghosh and Dubey, 2013) this research authors presented two important clustering algorithms. First is K-Means and second is an FCM (Fuzzy C-Means) algorithm and comparative study done between these algorithms.

In this (Jain and Gajbhiye, 2012) research authors presented competitive Neural Network, K means Clustering algorithm, Fuzzy C Means algorithm and one original method Quantum Clustering to consider its. The main aim of this research was to compare the four clustering techniques with lot capable presentation of sum data which is multivariate. Authors introduced

an easy-to-use and astute tool. This tool compares some clustering methods within the same scaffold in this research.

In (Kalyankar and Alaspurkar, 2013) this research authors presented a lot of amount of data mining concepts, data, many types of data mining methods such as Classification, clustering, hyper cyclic patterns, many types of algorithms etc.. The main aim of this research is study on weather data using data mining technique and methods like clustering technique.

In (Kavitha and Sarojamma, 2012) this research author presented Disease Management Programs are initial to encompass providers across the healthcare band. In this editorial creator express the diabetic monitoring platform, supports Diabetes (D) diagnosis assessment which is offering functions of DM. It is based on the CART method. This work explained a decision tree for diabetes diagnostics and showed how to use it as a basis for generating knowledge base rules.

In (Lavrač and Novak, 2013) this research worked on first outlines relational data mining approaches and finding of subgroups. In this research described recently developed approaches to semantic data mining which facilitate the use of sphere ontology. This ontology is the background acquaintance in analysis of data. The techniques and tools are useful of illustrated on selected biological applications.

The author in literature (Lai and Tsai, 2012) presented the preliminary results. This result present in the validation. The author presented risk measurement of total victory measures which upshot by profound hammering rains in the Shimen reservoir watershed of Taiwan. In this research author used spatial analysis and data mining algorithms. In this research researcher focused on Normalized Difference Vegetation Index, eleven factors such as Digital Elevation Model, DEM, slope, aspect, curvature, geology, soil, land use, fault, river and road.

In (Li and Ngom, 2013) this research worked on Tuning Non-negative matrix factorization is a matrix breakdown Non –negative matrix factorization is the create pattern such as matrix form. In this research author presents the data set in form of matrix decomposition.

In (Sujatha and Akila, 2012) this literature authors presented Breast Cancer prediction and diagnosis are two medical applications. It is great challenge to the researchers. This research summarizes various review and technical articles on diagnosis, breast cancer prediction presented an outline of the current research being passed out using the data mining techniques to improve the breast cancer prediction and diagnosis.

(Saha and Chaki , 2012) authors presented a brief review of protein sequence categorization. In this research authors involved the data mining applications. Data mining techniques and methods have been used by researchers for analyzing protein and DNA/RNA sequences. Researchers applied some well-known classification techniques like neural networks, Genetic algorithm, clustering, Fuzzy ARTMAP, Rough Set Classifier etc for correct categorization. Authors presented a review is with three different sorting models. First is fuzzy ARTMAP model second is neural network model and third is Rough set classifier model in this research.

In this (Sandhya et al. , 2013) research author presented a comparative analysis for future forecast of an shortened data set which used casual tree making and rough set in data mining. In this research the result found of simple classification technique is compared with the result of rough set attribute.

(Skurichina and Duin, 2001) presented Bagging and the random subspace method. RSM is the combining technique. Bagging works as a method of increasing accuracy. Bagging is also called Bootstrap. Bootstrap is based on random sampling with replacement. Bagging is a machine learning ensemble algorithm.

Methodology

This research deals with the methodological steps adopted in the research study. The researchProcedures followed are described under the following headlines:

A. Selection of Protein database and Forest fire database.

B. Selection of Data mining tools.

C. Use of Tools in research study such as MATLAB, WEKA.

D. Selection of Data mining Algorithms.

Reference

1. Agarwal,t.limielinski,and A.Swami mining association rules between sets of items in large databases,in ACM SIGMOD,1993.

2. Arthur David and Vassilvitskii Sergei (2007), "k-means++: The Advantages of Careful Seeding" ,SODA '07: Proceedings of the eighteenth annual ACM-SIAM,1027-1035.

3. Basheer M., Al-Maqaleh, Hamid Shahbazkia, (2012)," A Genetic Algorithm for Discovering Classification Rules in Data Mining", International Journal of Computer Applications (0975 – 8887), 41(18): 40-44.

4. Bifet A., Holmes G., Pfahringer B., Kirkby R., Gavald`a R.(2009), "New ensemble methods for evolving data streams", In *KDD*, ACM, 139-148.

5. Dietterich T. (2002), "Ensemble learning, in The Handbook of Brain Theory and Neural Networks", 2nd ed., M. Arbib, Ed., Cambridge MA: MIT Press.

6. Dr. Bhatia, M.P.S. and Khurana, Deepika (2013), "Experimental study of Data clustering using k-Means and modified algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP),3(3): 17-30.

7. Dr. Hemalatha M. and Saranya Naga N. (2011), "A Recent Survey on Knowledge Discovery in Spatial Data Mining" IJCSI International Journal of Computer Science, 8 (3):473-479.

8. Dzeroski S. and Zenko B. (2004), "Is combining classifiers with stacking better than selecting the best one? Machine Learning", 255–273.

9. Ester Martin, Kriegel Peter Hans, Sander Jörg (2001), "Algorithms and Applications for Spatial Data Mining Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS", Taylor and Francis, 1-32.

10. Elena Makhalova, (2013), "Fuzzy C means Clustering in MATLAB", The 7th International Days of Statistics and Economics, Prague, 19(21): 905-914.

11. Ester Martin, Kriegel Peter Hans, Sander Jörg (2001), "Algorithms and Applications for Spatial Data Mining Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS", Taylor and Francis, 1-32.

12. Ghosh, Soumi and Dubey, Sanjay, Kumar (2013), "Comparative Analysis of KMeans and Fuzzy C-Means Algorithms", (IJACSA) International Journal of Advanced Computer Science and Applications, 4(4) :35-39.

13. Jain Shreya and Gajbhiye Samta (2012), "A Comparative Performance Analysis of Clustering Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, 2(5):441-445.

14. Kalyankar A. Meghali and Prof. Alaspurkar S. J. (2013), "Data Mining Technique to Analyse the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering, 3(2):114-118.

15. Kavitha K. and Sarojamma R M (2012), "Monitoring of Diabetes with Data Mining via CART Method", International Journal of Emerging Technology and Advanced Engineering, 2 (11):157-162.

16. Lavrač Nada and Novak Kralj Petra (2013), "Relational and Semantic Data Mining for Biomedical Research", Informatica (Slovenia), 37(1): 35-39.

17. Lai J-S and Tsai F.(2012),"Verification And Risk Assessment For Landslides In The Shimen Reservoir Watershed Of Taiwan Using Spatial Analysis And Data Mining", International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXIX(B2): 67-70.

18. Li Yifeng, Ngom, Alioune (2013), "The Non-Negative Matrix Factorization Toolbox for Biological Data Mining", Source Code for Biology & Medicince, 8 (1): 1-18.

19. Mrs. Sujatha B. and Akila C., (2012), "Prediction of Breast Cancer Using Data Mining", International Journal of Computer Science and Management Research, 1 (3): 384-389.

20. Saha Suprativ and Chaki Rituparna,(2012)," A Brief Review of Data Mining Application Involving Protein Sequence Classification", International Journal of Database Management Systems (IJDMS),4(5).

21. Sandhya M., Dr. Kangaiammal A., Dr. Senthamarai C. (2013)," A Comparative Study on Decision Rule Induction for incomplete data using Rough Set and Random Tree Approaches", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727 9(3): 06-10.

22. Skurichina M. and Duin R.P.W. (2001) "Bagging and the random subspace method for redundant feature spaces", In: Kittler J, Roli F (eds) Multiple Classifier Systems (Proceedings Second International Workshop MCS 2001, Cambridge, UK). Lecture Notes in Computer Science, vol 2096. Springer-Verlag, Berlin, 1–10.